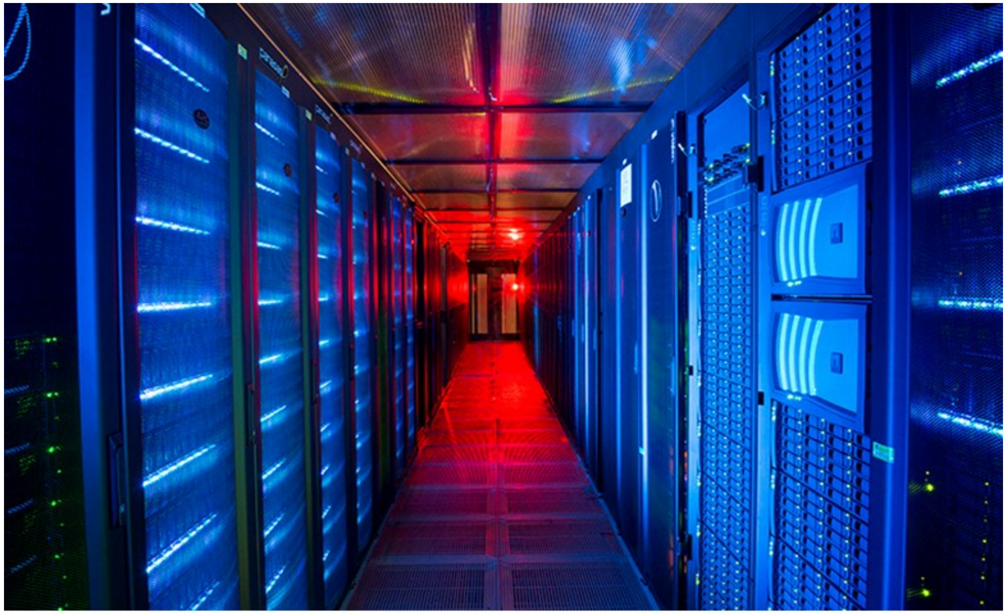
Science & Technology Facilities Council

JASMIN-CEMS: Big Data and Compute for Environmental Science



Victoria Bennett[1,3], Philip Kershaw[1,3],Matt Pritchard[1], Jonathan Churchill[2], Cristina Del Cano Novales[2], Martin Juckes[1,4], Stephen Pascoe[1,4], Sam Pepler[1,4], Ag Stephens[1,4], Bryan Lawrence[1,4,6]

Centre for Environmental Data Archival, RAL Space, STFC Rutherford Appleton Laboratory, UK; 2. Scientific Computing Department, STFC Rutherford Appleton Laboratory, UK; 3. National Centre for Earth Observation, UK; 4. National Centre for Atmospheric Science, UK; 5. Remote Sensing Group, RAL Space, STFC Rutherford Appleton Laboratory, UK; 6. University of Reading, UK

Victoria Bennett

CEDA, Centre for Environmental Data Archival, STFC

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Outline

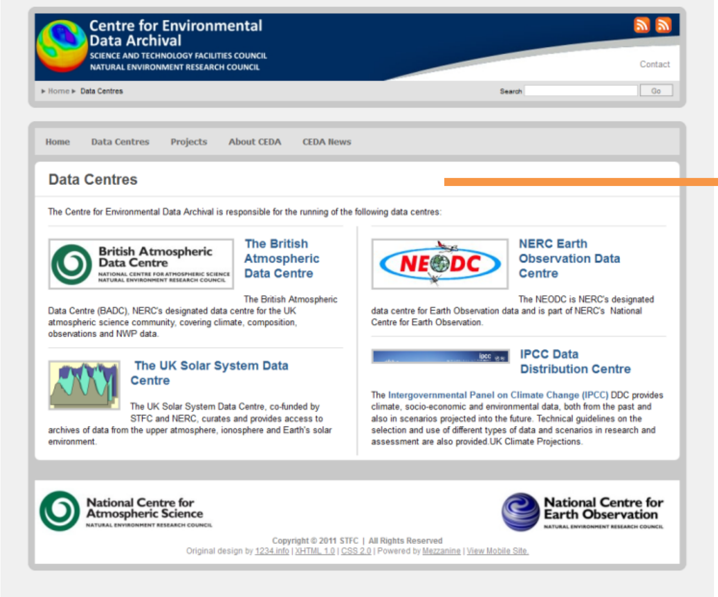
- Some background
- What is JASMIN, and CEMS
- Facts and figures
- JASMIN operations, and evolution
- Two example science projects

CEMS

Climate, Environment & Monitoring from Space

Background: CEDA

Centre for Environmental Data Archival



Centre for Environmental
Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Contact

▸ Home ▸ Data Centres

Search [ ] Go

Home | Data Centres | Projects | About CEDA | CEDA News

Data Centres

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

The British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

NERC Earth Observation Data Centre

The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.

The UK Solar System Data Centre

The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

IPCC Data Distribution Centre

The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.UK Climate Projections.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Copyright © 2011 STFC | All Rights Reserved
Original design by 1234.info | XHTML 1.0 | CSS 2.0 | Powered by Mezzanine | View Mobile Site.

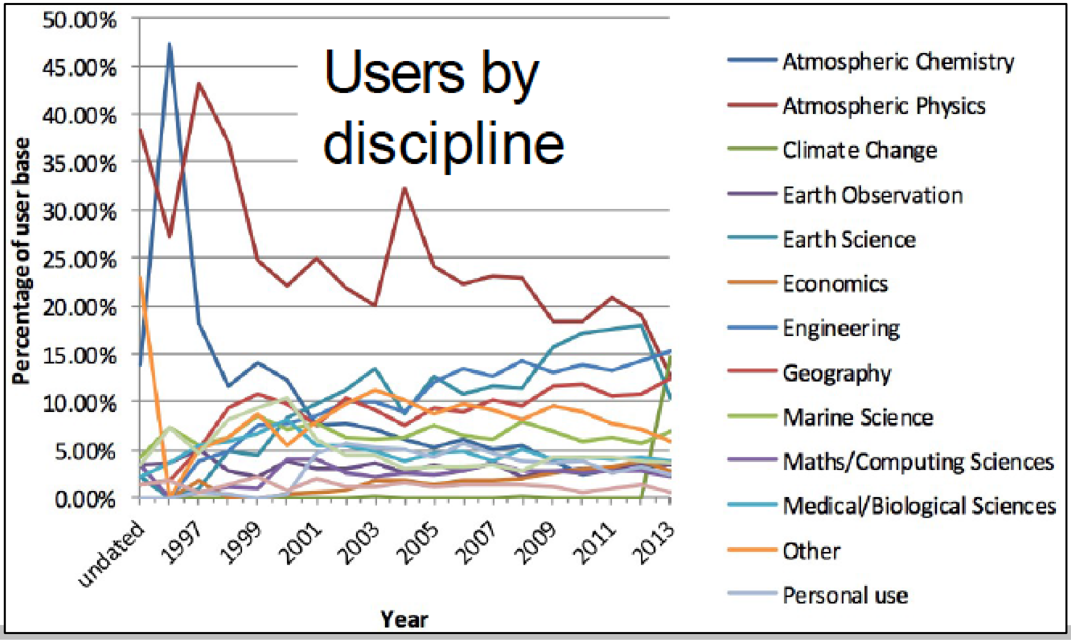
“to support environmental science, further environmental data archival practices, and develop and deploy new technologies to enhance access to data”

➔ Curation & Facilitation

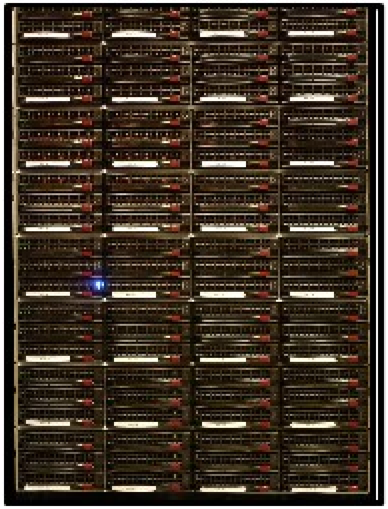
Centre for Environmental Data Archival

Project	Type	Data Volume (Petabytes)
NEODC	Earth Observation	0.9
BADC	Atmospheric Science	0.8
CMIP5	Climate Model	1.2
	Total	2.9

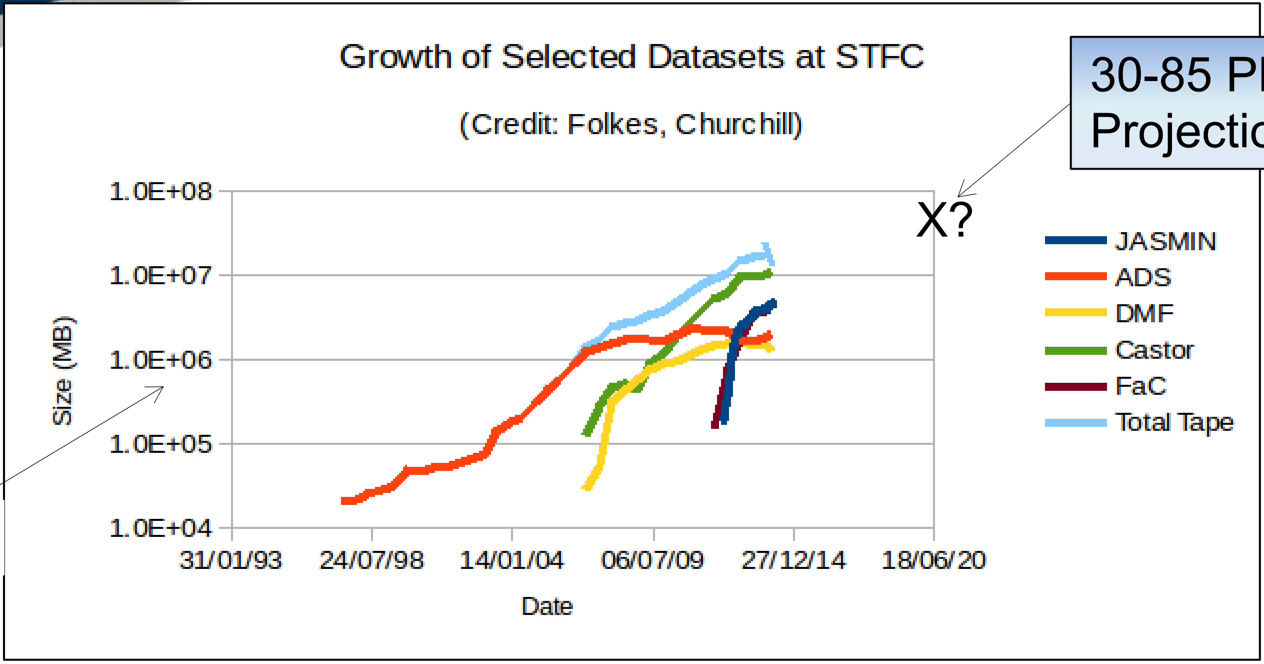
- > 300 datasets
- 144 million files
- 23,000 registered users



CEDA Evolution



Data growth at STFC



(Credit: Folkes, Churchill)

30-85 PB (unique data)
Projection for JASMIN

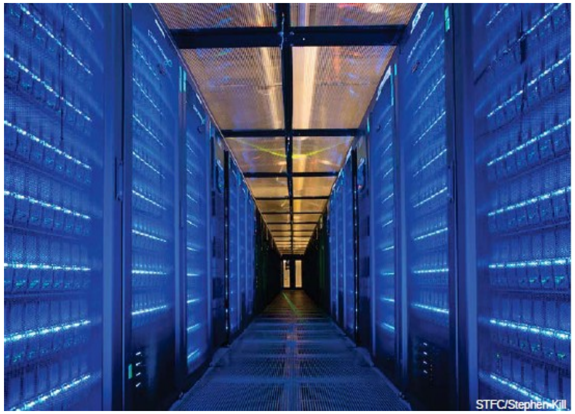
X?

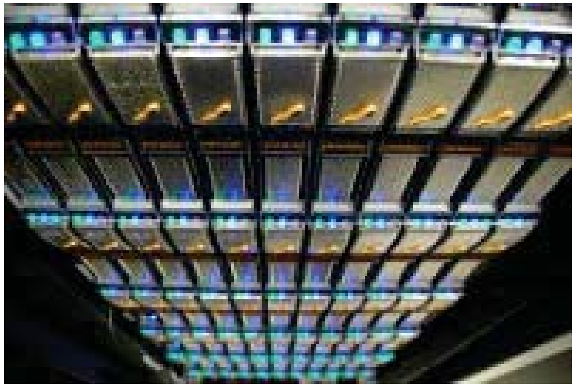
1 PB

CMIP6
30-300
PB?!

The light blue line is the total of all the data stored on tape in SCD.
The green line is the LHC Tier 1 data on tape.
The dark blue line is the data stored on **disk** in JASMIN.

JASMIN & CEMS: Big Data Facilities


STFC/Stephen Kill



- JASMIN (super data cluster)
  - Storage and services
  - Scientific computation
  - Access to high volume and complex data
- CEMS: Climate, Environment and Monitoring from Space
  - EO data and services
  - Academic – commercial partnership

CEMS

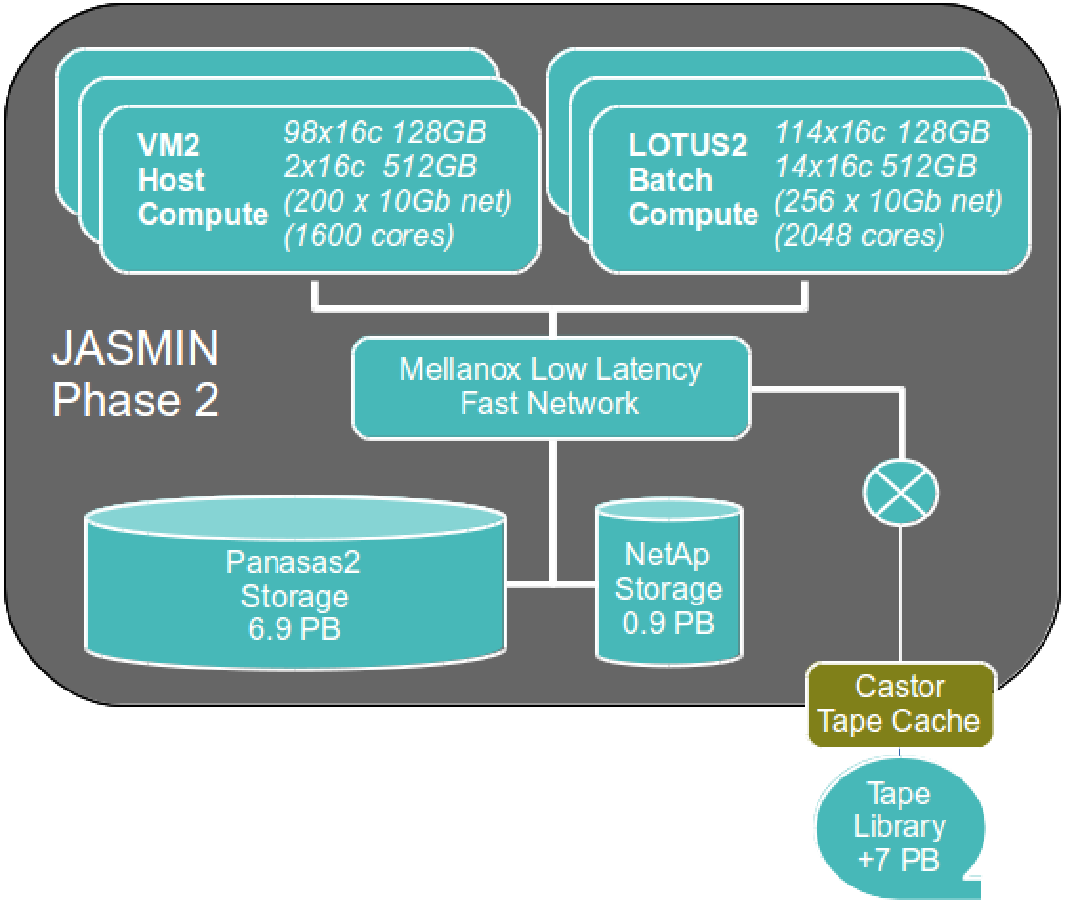
Climate, Environment & Monitoring from Space

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

CATAPULT
Satellite Applications

JASMIN

- What have we got:
  - ~16.4 PB fast parallel disk storage & equivalent in near-line tape
  - > 4,000 compute cores
- Four services provided to the community:
  - Storage (disk and tape)
  - Batch computing (“Lotus”)
  - Hosted computing
  - Cloud computing



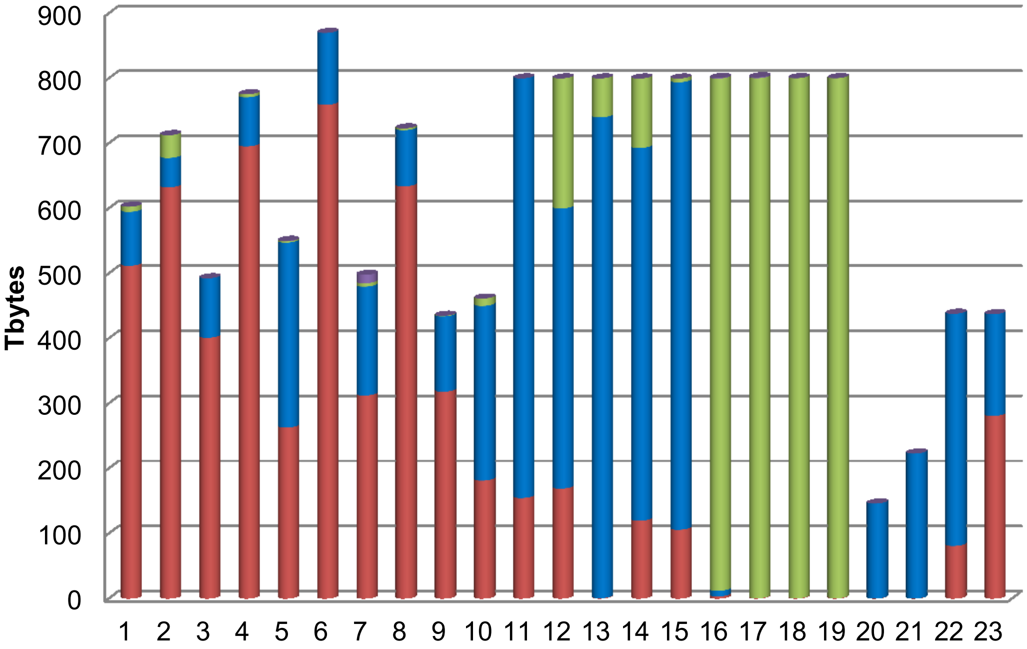
What are people using JASMIN-CEMS for?

- **Data collaboration** – share, process and disseminate;
- **Processing** – analyse own or third-party data;
- **Running models** – port, develop, share and run models;
- **Data/modelling services** – explore, develop and deploy services to provide new interfaces to end-users;
- **Cloud tools** – access to tools that allow creation of virtual servers and allocation of storage resources – for novel research/applications/tools

Activities with a research focus

JASMIN Operations

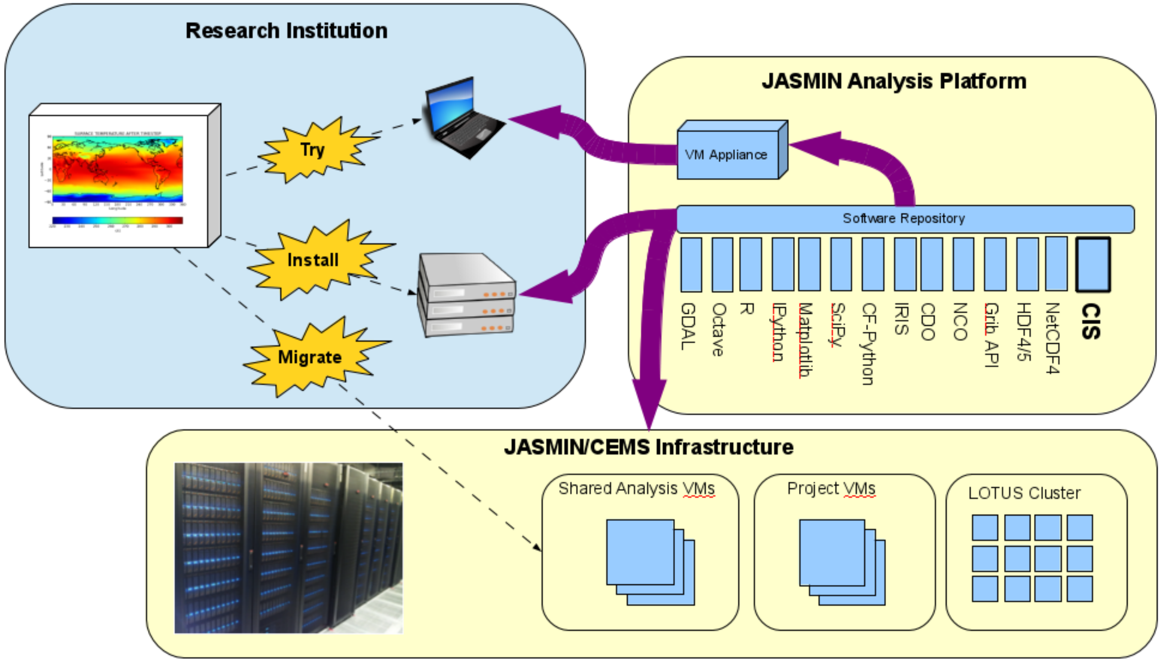
- ~600 JASMIN users
- 90 projects
- 5.2 PB allocated as Group Workspace; 3 PB CEDA archives
- Over 2 million processing jobs



JASMIN “bladesets” usage October 2014.
Blue: allocated but not yet used.
Red: used.
Green: as yet unallocated

JASMIN Analysis Platform (JAP)

- Multi-node infrastructure requires a way to install tools quickly and consistently
- The community needs a consistent platform where ever they need them.
- Users need help migrating analysis to JASMIN.
- JAP provides RPMs and pre-built images based on CentOS



http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/AnalysisPlatform

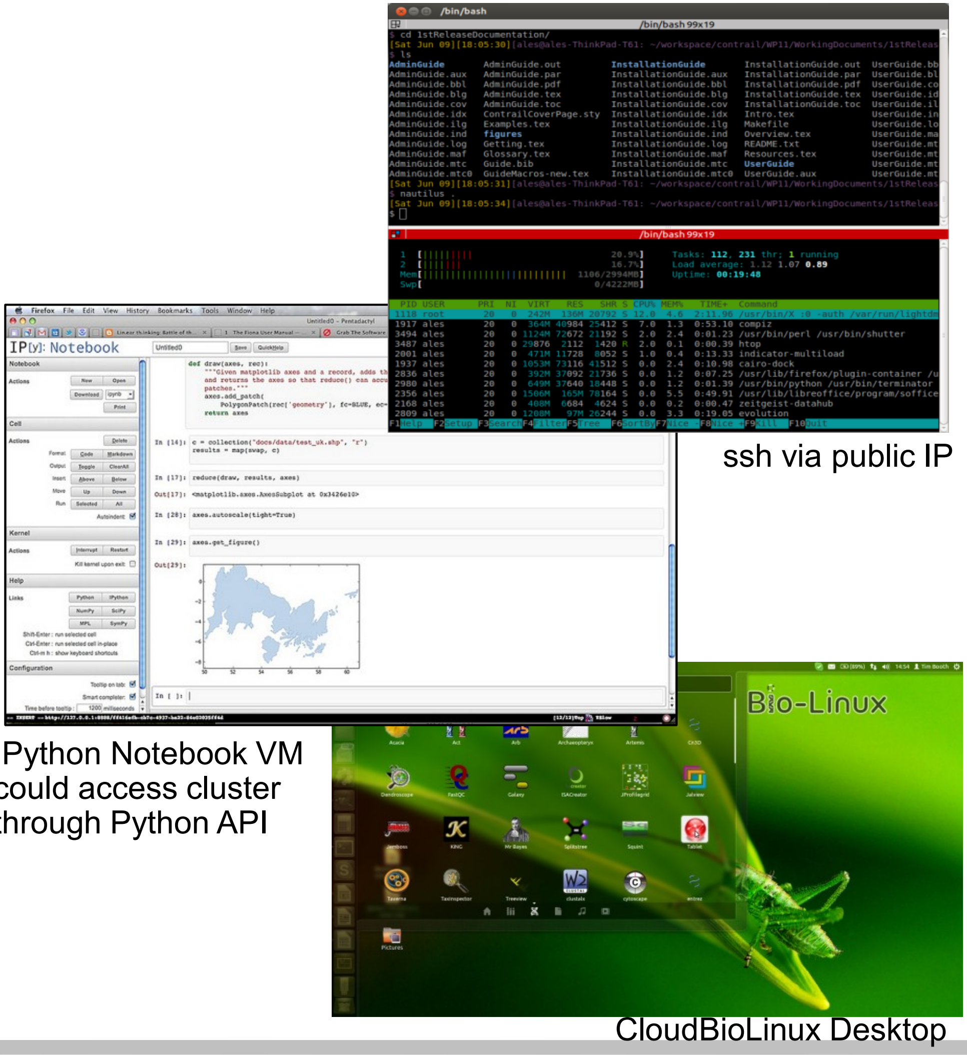
JASMIN Evolution

Phase 1:

Storage and batch compute – excellent results for first users..

But “long tail” of user community who are less expert users of e.g. the Linux command line and high performance computing

-> New cloud services to support much wider community



```
/bin/bash
/bin/bash 99x19
$ cd 1stReleaseDocumentation/
[Sat Jun 09][18:05:30][ales@ales-ThinkPad-T61: ~/workspace/contrail/WP11/WorkingDocuments/1stReleas
$ ls
AdminGuide       AdminGuide.out            InstallationGuide       InstallationGuide.out  UserGuide.bb
AdminGuide.aux   AdminGuide.par            InstallationGuide.aux   InstallationGuide.par  UserGuide.bl
AdminGuide.bbl   AdminGuide.pdf            InstallationGuide.bbl   InstallationGuide.pdf  UserGuide.co
AdminGuide.blg   AdminGuide.tex            InstallationGuide.blg   InstallationGuide.tex  UserGuide.id
AdminGuide.cov   AdminGuide.toc            InstallationGuide.cov   InstallationGuide.toc  UserGuide.il
AdminGuide.idx   ContrailCoverPage.sty     InstallationGuide.idx   Intro.tex              UserGuide.in
AdminGuide.ilg   Examples.tex              InstallationGuide.ilg   Makefile               UserGuide.lo
AdminGuide.ind   figures                   InstallationGuide.ind   Overview.tex           UserGuide.ma
AdminGuide.log   Getting.tex               InstallationGuide.log   README.txt             UserGuide.mt
AdminGuide.maf   Glossary.tex              InstallationGuide.maf   Resources.tex          UserGuide.mt
AdminGuide.mtc   Guide.bib                 InstallationGuide.mtc   UserGuide              UserGuide.mt
AdminGuide.mtc0  GuideMacros-new.tex       InstallationGuide.mtc0  UserGuide.aux          UserGuide.mt
[Sat Jun 09][18:05:31][ales@ales-ThinkPad-T61: ~/workspace/contrail/WP11/WorkingDocuments/1stReleas
$ nautilus .
[Sat Jun 09][18:05:34][ales@ales-ThinkPad-T61: ~/workspace/contrail/WP11/WorkingDocuments/1stReleas
$
```

```
/bin/bash 99x19
  1  [|||||||||                      20.9%]     Tasks: 112, 231 thr; 1 running
  2  [|||||||                        16.7%]     Load average: 1.12 1.07 0.89
  Mem[|||||||||||||||||||||||||  1106/2994MB]     Uptime: 00:19:48
  Swp[                                0/4222MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 1118 root       20   0  242M  136M 20792 S 12.0  4.6  2:11.96 /usr/bin/X :0 -auth /var/run/lightdm
 1917 ales       20   0  364M 40984 25412 S  7.0  1.3  0:53.10 compiz
 3494 ales       20   0 1124M 72672 21192 S  2.0  2.4  0:01.23 /usr/bin/perl /usr/bin/shutter
 3487 ales       20   0 29876  2112  1420 R  2.0  0.1  0:00.39 htop
 2001 ales       20   0  471M 11728  8052 S  1.0  0.4  0:13.33 indicator-multiload
 1937 ales       20   0 1053M 73116 41512 S  0.0  2.4  0:10.98 cairo-dock
 2836 ales       20   0  392M 37092 21736 S  0.0  1.2  0:07.25 /usr/lib/firefox/plugin-container /u
 2980 ales       20   0  649M 37640 18448 S  0.0  1.2  0:01.39 /usr/bin/python /usr/bin/terminator
 2356 ales       20   0 1506M  165M 78164 S  0.0  5.5  0:49.91 /usr/lib/libreoffice/program/soffice
 2168 ales       20   0  408M  6684  4624 S  0.0  0.2  0:00.47 zeitgeist-datahub
 2809 ales       20   0 1208M   97M 26244 S  0.0  3.3  0:19.05 evolution
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

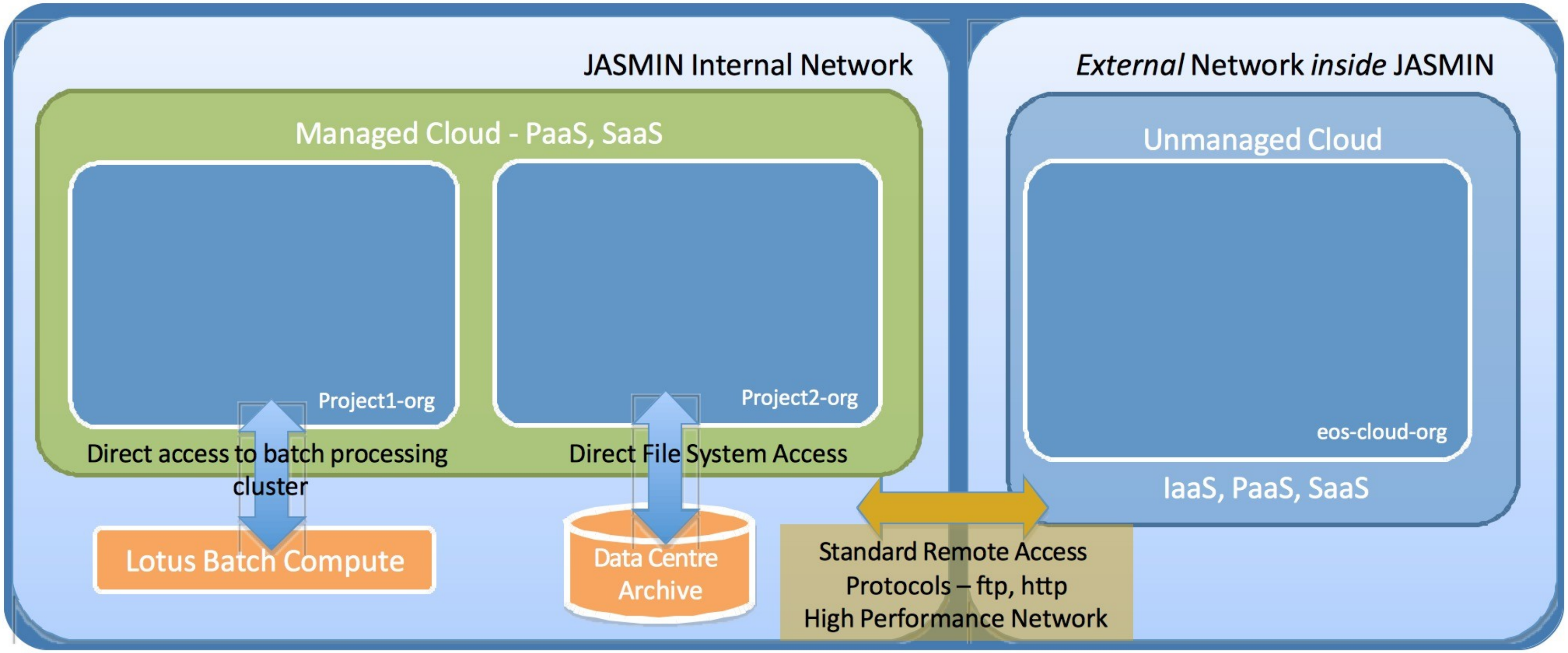
ssh via public IP

```
def draw(axes, rec):
    """Given matplotlib axes and a record, adds th
    and returns the axes so that reduce() can accu
    patches."""
    axes.add_patch(
        PolygonPatch(rec['geometry'], fc=BLUE, ec=
    return axes

In [14]: c = collection("docs/data/test_uk.shp", "r")
         results = map(swap, c)

In [17]: reduce(draw, results, axes)

Out[17]: <matplotlib.axes.AxesSubplot at 0x3426e10>

In [28]: axes.autoscale(tight=True)

In [29]: axes.get_figure()

Out[29]:

In [ ]:
```

IPython Notebook VM could access cluster through Python API

CloudBioLinux Desktop

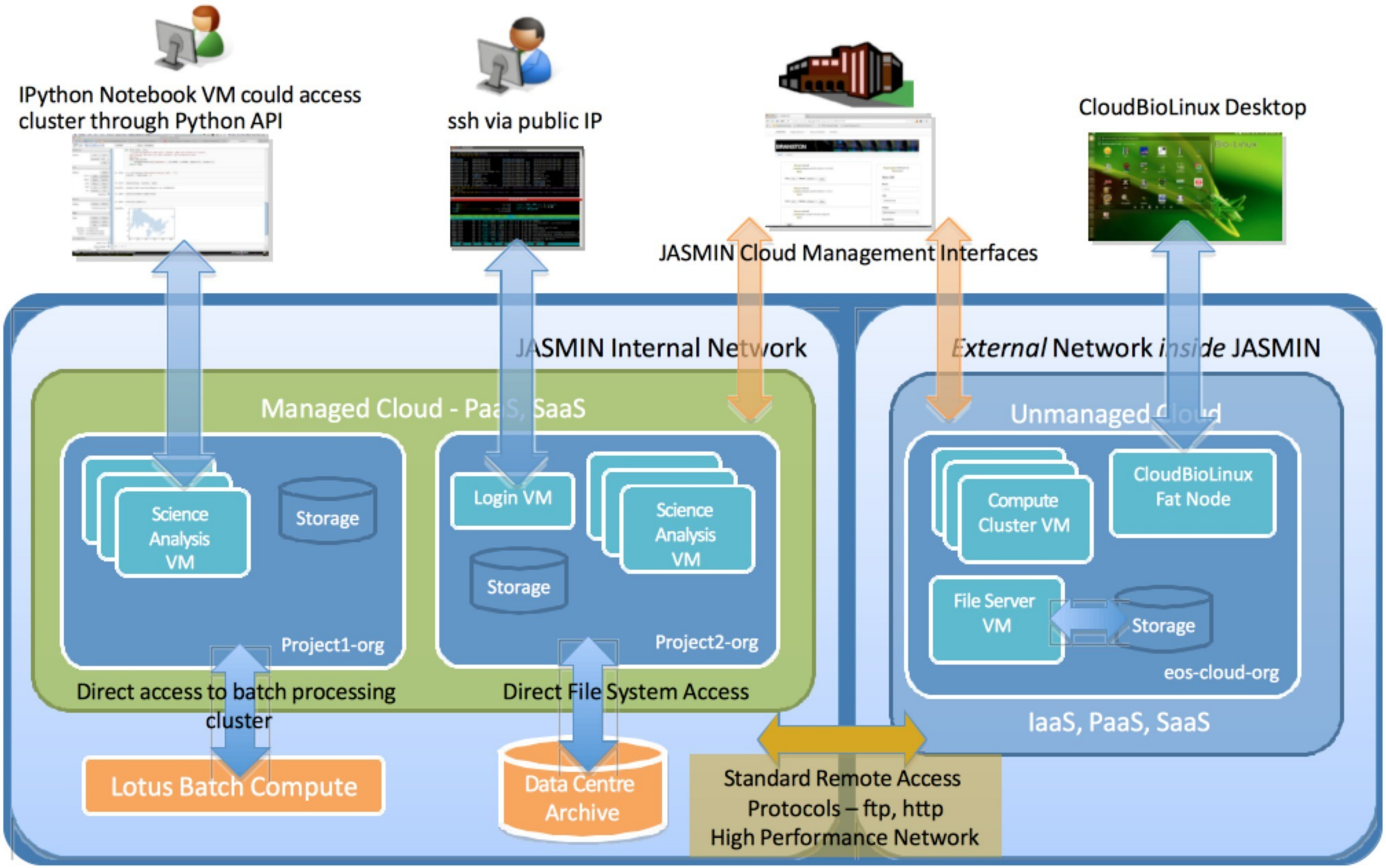
JASMIN Cloud Architecture


JASMIN Internal Network
Managed Cloud - PaaS, SaaS
Project1-org
Project2-org
Direct access to batch processing cluster
Direct File System Access
Lotus Batch Compute
Data Centre Archive
External Network inside JASMIN
Unmanaged Cloud
eos-cloud-org
IaaS, PaaS, SaaS
Standard Remote Access Protocols – ftp, http
High Performance Network

Group Work Spaces and hosted processing in the Managed Cloud:
direct access to archive filesystem and Lotus batch processing

First projects underway in the Unmanaged Cloud

JASMIN Cloud Architecture

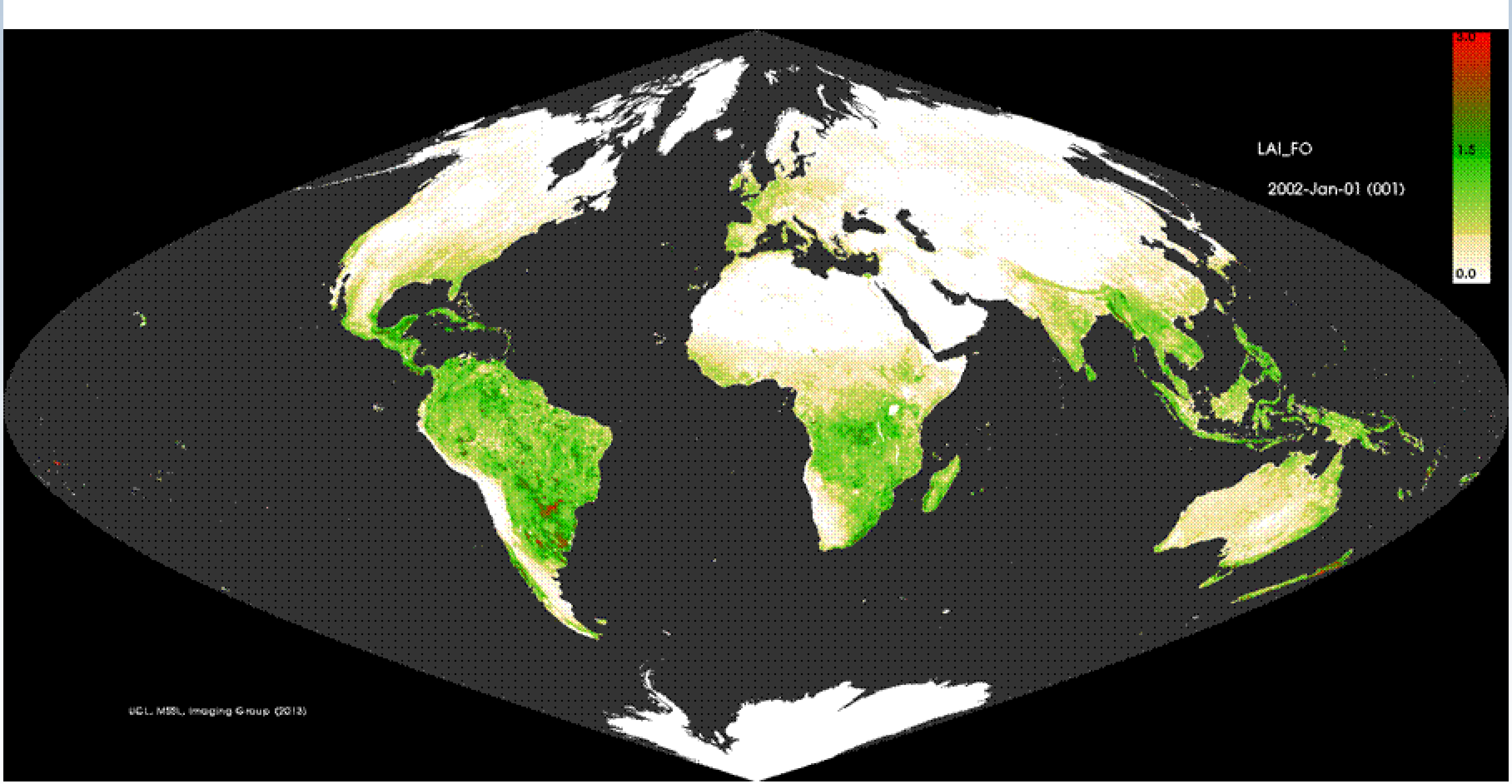


Example EO Science projects using JASMIN-CEMS

JASMIN-CEMS for global land surface products

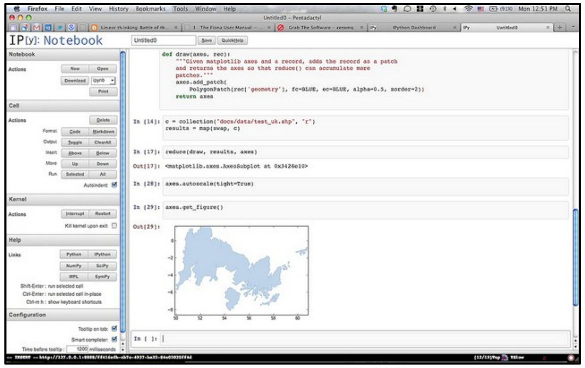
- **Objective 1:** Re-project BRDF files from SIN-coordinates to lat/lon
  - **Challenge**: huge number of polygons to be spatiality indexed and processed. **This process requires massive RAM and usually takes a very long time!**
- **Objective 2:** Create specific albedo products for computation of 8-daily LAI/fAPAR between 2002 and 2011 at 3 different resolutions: 1km, 5km and 25km
  - **Challenge:** Upscale big data BRDF (50TB) from 1km to 5km and 25km using energy conservation method: **This process is extremely time consuming!**
- **Solution**: Cloud-computing system in JASMIN-CEMS (~100 times faster than 224-core in house linux cluster)
- Also use Science DMZ for data transfers from NASA: achieved rates up to 28 TB/day

Example of 8-daily Global LAI derived from GlobAlbedo for 10 years using combined processing on CEMS and FastOpt Hamburg



ESA OPTIRAD Project

- Developing a Collaborative Research Environment for land data assimilation
  - a dedicated **software environment for the scientific community to generate products** from raw EO data
  - compute intensive assimilation algorithms with high memory demands
- Using iPython Notebook on the CEMS Unmanaged cloud



```
    def draw(axes, rec):
        """Given matplotlib axes and a record, adds the record as a patch
        and returns the axes so that reduce() can accumulate more
        patches."""
        axes.add_patch(
            PolygonPatch(rec['geometry'], fc=BLUE, ec=BLUE, alpha=0.5, zorder=2))
        return axes

In [14]: c = collection("docs/data/test_uk.shp", "r")
         results = map(swap, c)

In [17]: reduce(draw, results, axes)

Out[17]: <matplotlib.axes.AxesSubplot at 0x3426e10>

In [28]: axes.autoscale(tight=True)

In [29]: axes.get_figure()

Out[29]:

In [ ]:
```

Assimila UCL FastOpt vrije Universiteit amsterdam RAL Space

Further info

JASMIN

http://www.jasmin.ac.uk

Centre for Environmental Data Archival

http://www.ceda.ac.uk

JASMIN paper

Lawrence, B.N. , V.L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard, and A. Stephens. **Storing and manipulating environmental big data with JASMIN.** *Proceedings of IEEE Big Data 2013, p68-75,* doi:10.1109/BigData.2013.6691556

victoria.bennett@stfc.ac.uk